DOCUMENT RESUME

ED 435 705 TM 030 336

AUTHOR	Cook, Colleen
TITLE	A Review of Intraclass Correlation.
PUB DATE	2000-01-00
NOTE	35p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000).
PUB TYPE	Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Association (Psychology); *Behavioral Sciences; *Correlation; *Interrater Reliability; *Research; Research Design; *Sampling; *Statistical Significance
IDENTIFIERS	Confidence Intervals (Statistics); *Intraclass Correlation

ABSTRACT

Against an historical backdrop, this paper summarizes four uses of intraclass correlation of importance to contemporary researchers in the behavioral sciences. First, it shows how the intraclass correlation coefficient can be used to adjust confidence intervals for statistical significance testing when data are intracorrelated and the independence assumption is violated. Closely related to this discussion is a second application of the intraclass correlation coefficient to research design and sampling methodology in settings in which data are nonindependent. The third example is the use of the intraclass correlation coefficient as a measure of association. Finally, several versions of intraclass correlation coefficients, used as measures of interrater reliability among judges, are described. (Contains 7 tables, 2 figures, and 36 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 435 705

cook.wp1 11/11/99

A Review of Intraclass Correlation

Colleen Cook

Texas A&M University 77843-5000

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Colleen Cook
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM030336

BEST COPY AVAILABLE

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, January 27-29, 2000. The author may be contacted through e-mail address "ccook@tamu.edu".

Abstract

Against a historical backdrop, the present paper summarizes four uses of intraclass correlation of importance to contemporary researchers in the behavioral sciences. First, it is shown how the intraclass correlation coefficient can be used to adjust confidence intervals for statistical significance testing when data are intracorrelated and the independence assumption is violated. Closely related to this discussion is a second application of the intraclass correlation coefficient to research design and sampling methodology in settings in which data are non-independent. Third, the intraclass correlation coefficient as a measure of association is described. Fourth, several versions of intraclass correlation coefficients, used as measures of interrater reliability among judges are discussed.

The concept of intraclass correlation has a long history in statistics. In contrast to interclass correlation (e.g., the Pearson product-moment correlation--see Walsh [1996]), which evaluates the sameness in rank-orderings of data across classes or categories (e.g., variables), intraclass correlation relates to the sameness in rank-orderings of data within classes. Haggard (1958) defined the coefficient of intraclass correlation as "the measure of the relative homogeneity of the scores within the classes in relation to the total variation" (p. 6). The development of intraclass correlation was closely tied to the evolution of sophisticated statistical tools applied to problems in the behavioral sciences.

Against a historical backdrop, the present paper summarizes four uses of intraclass correlation of importance to contemporary researchers in the behavioral sciences. First, it will be shown how the intraclass correlation coefficient was used by Walsh (1947) to adjust confidence intervals for statistical significance testing when data are intracorrelated and the independence assumption is violated. Closely related to this discussion is a second application of the intraclass correlation coefficient to research design and sampling methodology in settings in which data are non-independent. Third, the intraclass correlation coefficient as a measure of association will be described. Fourth, several versions of intraclass correlation coefficients, used as measures of reliability among judges, as a special case of generalizability theory (cf. Kieffer, 1999; Shavelson & Webb, 1991; Thompson, 1991; Webb, Rowley & Shavelson, 1988), will be discussed.

Brief History

The intraclass correlation coefficient, rho, or ρ_I, was first developed to estimate family resemblance, e.g., the correlation in the height of brothers. As early as 1901 Pearson devised a method to compute a product-moment intraclass correlation coefficient from a symmetrical correlation table. A double entry was made for each pair of scores, thus for two brothers whose heights were 5'10" and 5'11", entries would be made for x=5'10", y=5'11" and x=5'11" and y=5'10". In this manner all possible correlations were computed at once, and the result was essentially the average of the correlations.

Because it is necessary to compute $k(k-1)$ entries for a k by k symmetrical table, the calculations rapidly become cumbersome (Haggard, 1958). Harris (1913), working with biological applications, devised a short cut for calculating the intraclass correlation coefficient based upon his discovery that, because variance between class means is separable from total variance, the intraclass correlation coefficient could be defined as the ratio of between-class variance to total variance.

Although Harris (1913) noted that intraclass correlation could be used as an effective analytical tool for problems in many fields (e.g., anthropology, sociology and related fields), behavioral scientists rarely used these research statistics through the first half of the 20th century. Intraclass correlation techniques, however, were used in applied science fields, such as agriculture and eugenics. For example, as a geneticist attempting to justify Darwin's theories in empirical terms, Fisher devoted extensive

coverage to the concept in both of his monumental monographs, Statistical methods for research workers (1925) and Design of experiments (1935); the former includes an entire chapter on the subject.

To Fisher is attributed the discovery that an unbiased estimate of the intraclass correlation coefficient could be explained in terms of mean squares in an analysis of variance (Haggard, 1958). Snedecor (1946) presented the formula in a convenient form adopted directly from an analysis of variance table (p. 245):

$$\rho_I = \frac{MS_B - MS_W}{MS_B + (n - 1)MS_W} \qquad (1)$$

where MS_B = Mean squares between--the F statistic numerator; MS_W = Mean squares within--the F statistic denominator; and n = Number of participants in a class.

In the numerator is the variance common to all individuals within a class. The denominator is an average variance for individuals if they had been chosen at random from the universe (Snedecor, 1946). The intraclass correlation coefficient measures the extent to which within group variability is small compared to between group variability. ρ_I is at maximum when data within groups are *identical* and means between groups are *different* (Shavelson, 1988).

In 1958, Haggard speculated that intraclass correlation was largely ignored by behavioral scientists, first because the fields of study were not yet sufficiently mature to utilize such statistical tools, and secondly, because for the first half of the

century researchers using statistical methods searched for the probability of the differences across classes or sample means rather than speculating on the sameness of data within classes.

Around the middle of the 20th century, the relevance of intraclass correlation to experimental design and sample survey research became apparent to behavioral scientists. It was finally recognized that score consistency within categories could be considered as nonindependence of scores within classes, and that this characterization could be an important red flag to researchers using statistical analyses (e.g., ANOVA, regression) presuming randomness of cases.

The concept of randomness is inherent in independence, which, with normality and homogeneity, form the three key assumptions underpinning analysis of variance testing. Long understood in its implications, the violation of the independence assumption has draconian effects upon the validity of statistical significance testing. Thus, Stevens (1986) noted that "the independence assumption is by far the most important assumption, for even a small violation of it produces a substantial effect on both the level of significance and the power of the F statistic" (p. 202).

Any close association of participants, such as exposure to a common classroom experience, or in the case of National Opinion Polling research, to a common neighborhood (Harris, 1997), can grossly exaggerate results. As Glass and Hopkins (1984) noted, "Randomization is important because it *helps* to ensure the independence of observations, or, equivalently, errors" (p. 350). But as Shaver (1993) explained, "Despite what is commonly assumed,

however, randomness does not guarantee independence. For example, observations almost certainly will not be independent when treatments have been delivered to subjects in a group setting, as is common in educational research" (p. 295).

Much research conducted by behavioral scientists is targeted at groups. One researcher may want to assess the effectiveness of a teaching method on classes of children; another may be conducting national opinion polls and may be constrained by cost to examining proximate geographic areas. The individuals with a common group membership, however, may affect the behavior, performance or opinions of one another. Individual student performance can be affected by the motivation or lack of motivation of other students in a given class. For example, a classroom continuously disrupted by an unruly student certainly constitutes a different experience for the students than they would have if they were members of a different class.

Similarly, Harris (1997) showed that people within neighborhoods in Great Britain tend to have similar political affinities; this would mean that randomly sampling individuals would not guarantee the independence of views of people sampled from the same neighborhood. It is oftentimes the case that an individual's experience in a group is not independent; rather there exists an intra-group exposure that must be considered when a study focuses on a class whose data are internally correlated and non-independent.

Significance Testing With Data Involving Intraclass Correlation

In 1947, Walsh investigated the effect of intraclass

correlation on confidence intervals and statistical significance tests (e.g., the t test of two means) derived under the assumption of independence. For example, classical analysis of variance operates under the assumption that scores within groups come from independent, identically distributed, normal, random variables. The ratio of the $\text{Mean Square}_{\text{BETWEEN}}$ to $\text{Mean Square}_{\text{WITHIN}}$ satisfies the F distribution iff (if and only if) the participants in the analysis have been independently sampled.

It has long been well understood that the existence of intraclass correlation could distort the analysis of variance in that the ratio of $\text{Mean Square}_{\text{BETWEEN}}$ to $\text{Mean Square}_{\text{WITHIN}}$ would be much larger. Because the $\text{sum-of-squares}_{\text{WITHIN}}$ (SS_{WITHIN}) is smaller when the data are not independent (i.e., there is intracorrelation), the denominator of the F ratio of these two variances is smaller than would be the case under a tenable assumption of independence.

Given assumptions of homogeneity and identical pairwise correlations, Walsh (1947) showed how to modify the ratio of SS_{BETWEEN} to SS_{WITHIN} to account for intraclass correlation in analysis of variance and produced a new statistic satisfying the F distribution. Importantly, from this result one can also determine how to modify confidence intervals in situations in which the independence assumption does not obtain as a result of the intraclass correlation effect.

Walsh (1947) computed tables showing how confidence intervals and statistical significance levels vary when samples are clustered and internally correlated, as contrasted with pure random samples. Walsh explained, "This shows that test results which appear to be

'significant' under the assumption of randomness are not necessarily 'significant' when [intraclass] correlation is present, even though the amount of correlation may be small" (p. 89).

Applying Walsh's constants, Barcikowski (1981) devised tables that dramatically showed how even a small amount of dependence among variables can cause actual α to be substantially greater than nominal α, where α is the testwise level of statistical significance, or the testwise probability of making a Type I error. *That is, because intraclass correlation makes a conventional* $F_{CALCULATED}$ *larger than it should be, intraclass correlation spuriously makes too many results statistically significant.*

For example, as illustrated in Figure 1, letting n denote group size, with nominal α of .05 and n=25 per cell, with an intraclass correlation coefficient of .05, actual α becomes .19; with n=50 per cell, and an intraclass correlation coefficient of .30, actual α becomes .68. Therefore, under conditions of nonindependence, the apparent statistical power of the test is exaggerated and the probability of committing a Type I error, i.e., rejecting the null when it is true, is greatly enhanced as the intraclass correlation among individuals and the numbers of participants in a group increase. Obviously, researchers should not conduct studies with so grossly inflated probabilities of committing Type I errors.

INSERT FIGURE 1 ABOUT HERE

Table 1 provides a heuristic data set for two groups of test scores to indicate the effect of intraclass correlation upon α.

Group 1 has a range of scores of 16, group 2 of 26. Table 2 reports the ANOVA summary table for these data.

INSERT TABLES 1 AND 2 ABOUT HERE

The intraclass correlation coefficient can be calculated for the Table 1 data using formula (1):

$$\rho_I = \frac{664.225 - 45.662}{664.225 + (19)(45.662)}$$

$$= \frac{618.563}{664.225 + 867.578}$$

$$= \frac{618.563}{1531.803}$$

$$= .404.$$

By consulting Figure 1 it can be estimated that an intraclass correlation coefficient of about .40 magnifies a nominal α of .05 to .60!

Basu, Odell and Lewis (1974) and Basu, Odell, Lewis and Kinderman (1975) built upon Walsh's findings in a univariate context and generalized non-independence issues to multiway environments. Smith and Lewis (1980) extended the discussion to a k-way (k-factor) experiment. Shoukri and Ward (1984) addressed the problem of unbalanced sample designs under non-independence. Donner (1986) and Olkin and Pratt (1958) reviewed theory and methodology for inferences using the intraclass correlation coefficient.

Intraclass Correlation and Sampling Methodology

When a treatment is administered to an individual, the proper unit of study is the *individual*, but in many applications in the

behavioral sciences, where treatment is administered to groups involving interaction among the individuals (e.g., in classrooms or therapy groups), the proper units of analysis are the *groups*. In influencing one another, the individuals create a group dynamic that is nonindependent and intracorrelated. The validity of a study is compromised when treatments are given to groups and statistical analyses are performed with individuals' data.

In his classic work, Statistical analysis in educational research, Lindquist (1940) discussed ANOVA using group means instead of individuals as units of analysis to avoid the dependence issue inherent in much group-focused inquiry. Barcikowski (1981) noted that most researchers through the 1960's ignored Lindquist's recommendation to focus on group means, based on the real concern that reducing n to the number of group means would seriously degrade the statistical power of the test to detect a statistically significant result.

However, using group means does not inherently result in as dramatic a loss in power as first thought, because the sets of means have less variability than do the corresponding sets of individual scores (Stevens, 1986, p. 204). That is, for a fixed number of degrees of freedom and a fixed explained sum of squares, smaller within-groups variation will itself result in a smaller calculated F. Barcikowski (1981, p. 268) determined how many subgroups there must be to detect meaningful differences between treatment groups when independent group means, obtained by averaging over related scores, are used as units of analysis. He presented tables that indicate the power of F tests when group

means are used as units of analysis under various conditions.

Figure 2 portrays some illustrative results regarding the use of group means as the unit of analysis. The figure presents, for group sizes ranging from 10 to 40, the minimum number of *groups* necessary for statistical power against Type II error to be greater than .8 in a one-way two-treatment (i.e., the *groups* are assigned to either of the two treatments) design for effect sizes of ¦.10¦, ¦.25¦, and ¦.40¦ for $\alpha = .05$.

INSERT FIGURE 2 ABOUT HERE

By consulting the tables provided by Barcikowski (1981) and reprinted by Stevens (1986), a researcher may determine in advance the number of entities necessary in a cell to obtain a given power and effect size. For instance, if the population intraclass correlation coefficient is .05 and α is .01 with an effect size of .10, then 572 independent "groups" each involving one participant would be necessary for a power of .8, or 85 groups with 10 participants per group would be needed, or 67 groups with 15 participants per group would be needed (Barcikowski, 1981, p. 279).

Barcikowski (1981) drew several conclusions of interest to researchers. The number of groups necessary to attain reasonable power ($\geq$ to .80) is dependent on the level of statistical significance, the correlation within groups, the expected effect size and the group size. First, under conditions of group means as units of analysis and an intraclass correlation effect, more participants are needed in an ANOVA design to attain the same power than under conditions of independence. Second, power declines as

the relatedness of participants in groups increases; as a result more groups are needed to attain acceptable power levels of .80 or greater. Third, as group size increases, more participants are needed in a cell to maintain a given power level. And finally, power can drastically decline when a small number of groups is being used and one or several groups are lost.

Intraclass Correlation as a Strength of Association Measure

A third major use of the intraclass correlation coefficient is as a strength of association measure. In that a statistically significant F test does not yield information regarding the magnitude of a treatment effect (cf. Cohen, 1994; Kirk, 1996; Thompson, 1996), the calculation of an appropriate strength of association measure is encouraged to provide insight into the potency of an effect. As the APA Task Force on Statistical Inference recently emphasized, "Always provide some effect-size estimate when reporting a p value" (Wilkinson & The APA Task Force on Statistical Inference, 1999, p. 599, emphasis added). Later the Task Force also wrote,

> Always present effect sizes for primary outcomes.... We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. (p. 599, emphasis added)

The intraclass correlation coefficient, calculated through formula (1), is the proper measure of association index in a random-effects ANOVA environment (see Frederick, 1999). The ratio provides an indication of the proportion of variance in the

dependent variable resulting from the independent variable dynamic and gauges the extent of within-group variability compared to between-group variability. ρ_I is 1.00 when within-group scores are identical and between group means vary (Shavelson, 1988, p. 363).

Intraclass Correlation as an Index of Interrater Reliability

The literature discusses a fourth use of intraclass correlation, as a measure of interrater reliability. In the 1950's and 1960's, intraclass correlation formulas were adapted for use in reliability theory. Ebel (1951), Haggard (1958), and Buros (1963) showed how intraclass correlation coefficients could be used as reliability statistics. Cronbach, Gleser, Nanda and Rajaratnam (1972) incorporated intraclass correlation into their generalizability theory.

The question of reliability of results is readily apparent in the case of judges rating a group of targets. For instance, a researcher may wish to gain insight into consistency in grades assigned by a set of instructors to a group of student papers, or ascertain consensus among a panel of referees who evaluate publication potential of a set of manuscripts. The question of consistency and reliability of ratings among judges is critical to research in the behavioral sciences; assessing whether observers agree or disagree on judgments of personalities or grading is critical in analyzing the integrity of results.

Shrout (1995) issued a caveat, however; in any study of consensus, it is important to keep in mind that consensus is not synonymous with accuracy. Intraclass correlation coefficients may be useful in assessing the consistency in ratings among several

judges; they do not shed light on the accuracy of the judgements--raters might agree among themselves, and still all be wrong (pp. 81-82).

As a special case of generalizability theory, Shrout and Fleiss (1979) described six forms of intraclass correlation coefficients used to measure interrater consensus under random and fixed conditions. To choose the correct form of intraclass correlation three questions must be considered: (a) is a one- or two-way ANOVA appropriate for the study?; (b) are differences in the mean ratings of judges of relevant research interest?; and (c) is the unit of analysis the individual judge or the mean of judges? (p. 420). In most studies, each of a random sample of n targets is judged independently by k raters. Under such conditions, Shrout and Fleiss (1979) describe three cases:

Case 1: each target is judged by a different set of raters randomly chosen from a larger population of judges (i.e., sampling with replacement);

Case 2: a random sample of k raters is selected from a larger population and each rater judges each of n targets (i.e., sampling without replacement); or

Case 3: each target is judged by each of the same k judges and those are the only judges of research interest (p. 421).

Each of the three cases requires a different mathematical model for computing an intraclass correlation coefficient based upon different assumptions and different ANOVA constructions. When the intraclass correlation approaches one, interrater consensus is high; when residual variance is high relative to

target variance, consensus is low, and the intraclass correlation coefficient will approach zero (Shrout, 1995, p. 85).

Case 1

Variability in Case 1 due to judges, target, judge/target interaction and random error is not separable. The appropriate model for analysis is a one-way random effects ANOVA and yields a between-targets measure and a within-targets measure. Under a random-effects model, results can be generalized to the population from which the independent variable was selected (Hinkle, Wiersma & Jurs, 1998, p. 449). The intraclass correlation coefficient, ICC (1,1), is calculated by formula (1) (Shrout & Fleiss, 1979, p. 421).

Case 2

Variability in Case 2 is further subdivided and a target-by-judges two-way random-effects model ANOVA is the appropriate model and, as a result, data regarding judges can be generalized to other judges in the population. The between-targets measure is maintained, while the within-target SS is partitioned into a between-judges SS and an error SS (Shrout & Fleiss, 1979, pp. 421-422). The intraclass correlation coefficient, ICC (2,1), is calculated by the following formula attributed to Rajaratnam (1960) and Bartko (1966):

$$ICC\ (2,1) = \frac{TMS - EMS}{TMS + (k-1)EMS + k(JMS-EMS)/n}$$

where TMS = target mean squares from the two-way ANOVA table, EMS = error mean squares from the two-way ANOVA table, and JMS = judge mean squares from the two-way ANOVA table.

Case 3

Case 3 is similar to Case 2 in that variability in Case 3 is also subdivided and a target-by-judge two-way ANOVA is the appropriate model. Variability is broken into a between-targets measure, and within-targets SS which is subdivided into a between-judges SS and an error SS. Case 3 differs from Case 2 in that it operates under a mixed-effects model (see Frederick, 1999).

Randomness is assumed for targets and the interaction between targets and judges, but the main effect, judges, is fixed. That is, interest is in a fixed set of judges with no consideration of the population from which they were derived. In a mixed effects model, randomness is assumed for the independent main effect and the interaction, while the other main effect is considered fixed. The intraclass correlation coefficient, ICC (3,1), under Case 3 is calculated as follows (Shrout & Fleiss, 1979, p. 422):

$$\text{ICC } (3,1) = \frac{\text{TMS} - \text{EMS}}{\text{TMS} + (k-1)\ \text{EMS}}.$$

Issues in Case Selection

In choosing among the various forms of the intraclass correlation coefficient outlined by Shrout and Fleiss (1979), two issues must first be addressed: experimental design and the conceptual intent of the study. To address the issue of experimental design one must decide whether a one- or two-way ANOVA is appropriate. Shrout and Fleiss noted that it is unlikely that ICC (2,1) or ICC (3,1) would be incorrectly applied in a Case 1, ICC (1,1) situation. The reverse is not true, however. It is quite likely that ICC (1, 1) might erroneously be applied to Case 2 and

Case 3 situations, resulting in an underestimate of ρ_I (p. 422).

A second important decision relates to the conceptual intent of the study (i.e., are effects due to judges important to the reliability index?). In choosing whether to apply Case 2 or Case 3, one is, in essence, deciding whether the judges represent random (i.e., Case 2) or fixed (i.e., Case 3) effects. The consistency of ratings is measured by ICC (3,1) if judges rate the same n targets, because the judges are considered fixed effects. If the consideration is whether the judges are interchangeable, ICC (2,1) is appropriate because the judges are considered random effects (Shrout & Fleiss, 1979, p. 425).

Yet a third important consideration is the appropriate unit of analysis, i.e., the individual judge or the mean of judges. Shrout and Fleiss (1979) suggested that it is a rare occurrence when a mean rating is substantively used as the unit of analysis. The mean is used when the individual ratings are unreliable (p. 426). Versions of the intraclass correlation coefficient that are relevant to consensus use the individual judge ratings, not the mean (Shrout, 1995, p. 85). Shrout and Fleiss provided modifications of the ICC formulas necessary when the mean, rather the individual judge, is the focus of study for each of the three cases. Table 3 outlines all six Shrout and Fleiss intraclass correlation coefficients and conditions for their application.

INSERT TABLE 3 ABOUT HERE

Factors Biasing Estimates

Statistical inference for consensus requires that ratings be

normally distributed. Under conditions of normality, F tests the null that there is no consensus among judges (Shrout, 1995, p. 87).

Shrout (1995) noted limitations of using ANOVA to measure reliability. First, since ANOVA focuses on absolute rating levels, and assumes that the variance of ratings is the same across judges and targets, judge elevation, or bias, may skew results. Second, another caveat arises when judges even implicitly or unconsciously use different evaluation scales (e.g., some judges with less information regarding the targets might employ a smaller range of scores than judges with more in-depth knowledge of targets). Thus, variance in ratings would differ among judges.

Third, target variation influences consensus. When error is fixed, the intraclass correlation coefficient is reduced as target variation is reduced. High consensus may result from heterogeneous target populations, and low consensus in homogeneous populations. If target variation were small, the intraclass correlation would approach zero, even if judge consensus was high (p. 90).

Heuristic Comparisons

For comparison purposes, all six formulae presented in Table 3 are applied to the data set presented in Table 4. Table 5 presents the ANOVA summary table for these data. Table 6 presents the six sets of calculations. Table 7 presents a comparison of the six estimates for the Table 4 data.

INSERT TABLES 4 THRU 7 ABOUT HERE

Conclusions

Applied not long after the turn of the century by Harris (1913) and Fisher (1925, 1935) to biology and genetics research, the concept of the intraclass correlation is as old as the field of statistics itself. There are many venues for applying the intraclass correlation coefficient, four of which have been discussed here.

First, intraclass correlation is important to the assumption of independence in several ways. When data are correlated within class, Walsh (1947) has shown how the intraclass correlation coefficient may be used to adjust confidence intervals by a constant and still maintain an F distribution to permit statistical significance testing with classical analyses (e.g., ANOVA, t test). Using Walsh's formulas, Barcikowski (1981) created tables indicating statistical significance levels, effect sizes and power levels under various intraclass correlation coefficients.

It is noteworthy, notwithstanding the exhortations to researchers sprinkled throughout the literature, that violating independence assumptions has draconian effects on classical statistical results, and such violations are all too commonly encountered. The lessons of nonindependence are ones that researchers seem to lose over time, and rediscover in keeping with the old adage that those who forget the past are condemned to repeat it.

The literature also shows that intraclass correlation has been used as a measure of association in random effects ANOVA and as a reliability measure, notably in generalizability theory as explained by Cronbach et al. (1972). Shrout and Fleiss (1979)

presented a special case of generalizability theory in guidelines for the use of six forms of intraclass correlation coefficients for evaluating the interrater reliability of judges.

References

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. Journal of Educational Statistics, 6, 267-285.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. Psychological Bulletin, 83, 762-765.

Basu, J. P., Odell, P. L. & Lewis, T. O. (1974). The effects of intraclass correlation on certain significance tests when sampling from multivariate normal populations. Communications in Statistics, 9, 899-908.

Basu, J. P., Odell, P. L., Lewis, T. O. & Kinderman, A. (1975). On independence of sample mean and translation invariant statistics of samples from multivariate normal populations. Journal of the American Statistical Association, 70, 480-481.

Buros, O. K. (1963). Schematization of old and new concepts of test reliability based upon parametric models. New Brunswick, NJ: Gryphon Press.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). The dependability of behavioral measurement. New York: John Wiley.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. International Statistics Review, 54 (1), 67-82.

Ebel, R. L. (1951). Estimation of the reliability of ratings. Psychometrika, 16, 407-424.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935). The design of experiments. Edinburgh: Oliver and Boyd.

Frederick, B.N. (1999). Fixed-, random-, and mixed-effects ANOVA models: A user-friendly guide for increasing the generalizability of ANOVA results. In B. Thompson (Ed.), Advances in social science methodology (Vol. 5, pp. 111-121). Stamford, CT: JAI Press.

Glass, G. V, & Hopkins, K. D. (1984). Statistical methods in education and psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Haggard, E. A. (1958). Intraclass correlation and the analysis of variance. New York: The Dryden Press.

Harris, J. A. (1913). On the calculation of the intraclass and interclass coefficients of correlation from class moments when the number of possible combinations is large. Biometrika, 9, 446-472.

Harris, P. (1997). The effect of clustering on costs and sampling errors of random samples. Journal of the Market Research Society, 39 (1), 41-51. (Original work published 1977).

Hinkle, D. E., Wiersma, W. & Jurs, S. G. (1998). Applied statistics for the behavioral sciences (4th ed.). Boston: Houghton Mifflin.

Kieffer, K.M. (1999). Why Generalizability Theory is essential and classical test theory is often inadequate. In B. Thompson (Ed.), Advances in social science methodology (Vol. 5, pp. 149-

170). Stamford, CT: JAI Press.

Kirk, R. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Lindquist, E. (1940). Statistical analysis in educational research. Boston: Houghton Mifflin.

Olkin, I., & Pratt, J.W. (1958). Unbiased estimation of certain correlation coefficients. Annals of Mathematical Statistics, 29, 201-211.

Shavelson, R. J. (1988). Statistical reasoning for the behavioral sciences. Boston: Allyn and Bacon.

Shavelson, R., & Webb, N. (1991). Generalizability theory: A primer. Newbury Park, CA: SAGE.

Shaver, J. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61, 293-316.

Shoukri, M. M. & Ward, R. H. (1984). On the estimation of the intraclass correlation. Communications in Statistics, 13, 1239-1255.

Shrout, P. E. (1995). Measuring the degree of consensus in personality judgments. In P. C. Shrout & S. T. Fiske (Eds.), Personality, research, methods, and theory (pp. 79-92). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Smith, J. H. & Lewis, T. O. (1980). Determining the effects of intraclass correlation on factorial experiments. Communications in Statistics, 13, 1353-1364.

Snedecor, G. W. (1946). Statistical methods. Ames, IA: Iowa State College Press.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum.

Thompson, B. (1991). Review of Generalizability theory: A primer by R.J. Shavelson & N.W. Webb. Educational and Psychological Measurement, 51, 1069-1075.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Walsh, B.D. (1996). A note on factors that attenuate the correlation coefficient and its analogs. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 21-32). Greenwich, CT: JAI Press.

Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. Annals of Mathematical Statistics, 18, 88-96.

Webb, N., Rowley, G., & Shavelson, R. (1988). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, 81-90.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604. [reprint available through the APA Web Page: **http://www.apa.org/journals/amp/amp548594.html**]

Table 1
An Example of Nominal vs. Actual alpha under Non-Independence

Class #1		Class #2	
ID	Score	ID	Score
1	82	21	65
2	91	22	67
3	86	23	83
4	88	24	77
5	93	25	90
6	94	26	65
7	82	27	89
8	80	28	81
9	95	29	73
10	82	30	81
11	91	31	90
12	88	32	87
13	87	33	80
14	85	34	79
15	92	35	85
16	93	36	88
17	85	37	77
18	88	38	89
19	89	39	68
20	83	40	77

Note. In class #1 scores range from 80 to 95; in class #2 scores range from 65 to 90.

Table 2
Analysis of Variance for Exam Scores

Source	Sum of Squares	df	Mean Squares	F_{CALC}	p_{CALC}
Between Classes	664.225	1	664.225	14.547	.00049
Within Classes	1735.150	38	45.662		
Total	2399.375	39			

Note. Although SPSS reports p_{CALC} to be ".000", meaning that an impossible sample result (p = .000) has been obtained, which is impossible, the correct p value can be found by evaluating the F_{CALC} using the spreadsheet function, "=fdist(15.547, 1, 38)".

Table 3
Intraclass Correlation Coefficients and Conditions for Their Use

Intraclass Correlation Coefficient	Formula	ANOVA	Model	Unit of Analysis
ICC (1,1)	$\frac{BMS-WMS}{BMS+(k-1)WMS}$	One-way	Random targets Random judges	Individual
ICC (2,1)	$\frac{BMS-EMS}{BMS+(k-1)EMS+k(JMS-EMS)/n}$	Two-way	Random targets Random judges	Individual
ICC (3,1)	$\frac{BMS-EMS}{BMS+(k-1)EMS}$	Two-way	Random targets Fixed judges	Individual
ICC (1,k)	$\frac{BMS-WMS}{BMS}$	One-way	Random targets Random judges	Mean
ICC (2,k)	$\frac{BMS-EMS}{BMS+(JMS-EMS)/n}$	Two-way	Random targets Random judges	Mean
ICC (3,k)	$\frac{BMS-EMS}{BMS}$	Two-way	Random targets Fixed judges	Mean

Note. n = number of Targets; k = number of Judges. Summarized from Shrout and Fleiss (1979).

Table 4
Three Ratings on Five Targets

Target	Judge		
	1	2	3
1	7	4	4
2	5	4	3
3	7	5	4
4	2	3	1
5	7	5	2

Table 5
Analysis of Variance for Ratings

Source of Variation	df	Mean Square
Between Targets	4	5.267
Within Target	10	2.733
Between Judges	2	9.800
Residual	8	.967

Table 6
Six Intraclass Coefficients for Table 4 Data Set

ICC Situation	ICC Situation
ICC (1,1)	ICC (1,k) or (1,3)
$\frac{BMS - WMS}{BMS + (k-1)(WMS)}$	$\frac{BMS - WMS}{BMS}$
$\frac{5.267 - 2.733}{5.267 + (2)(2.733)}$	$\frac{5.267 - 2.733}{5.267}$
$\frac{2.534}{5.267 + 5.466}$	$\frac{2.534}{5.267}$
$\frac{2.534}{10.733}$	**.481**
.236	
ICC (2,1)	ICC (2,k) or (2,3)
$\frac{BMS - EMS}{BMS+(k-1)EMS+k(JMS-EMS)/n}$	$\frac{BMS - EMS}{BMS+(JMS-EMS)/n}$
$\frac{5.267 - .967}{5.267 + (2)0.967 + (3)(9.8-.967)/5}$	$\frac{5.267 - .967}{5.267 - (9.8 - .967)/5}$
$\frac{4.3}{5.267 + 1.934 + (3)8.833/5}$	$\frac{4.3}{5.267 - 8.833/5}$
$\frac{4.3}{5.267 + 1.934 + 26.499/5}$	$\frac{4.3}{5.267 - 1.7666}$
$\frac{4.3}{5.267 + 1.934 + 5.2998}$	$\frac{4.3}{7.0336}$
$\frac{4.3}{12.5008}$	**.611**
.344	
ICC (3,1)	ICC (3,k)
$\frac{BMS - EMS}{BMS + (k-1)(EMS)}$	$\frac{BMS - EMS}{BMS}$

$$\frac{5.267 - .967}{5.267 + (2).967}$$

$$\frac{4.3}{5.267 + 1.934}$$

$$\frac{4.3}{7.201}$$

.597

$$\frac{5.267 - .967}{5.267}$$

$$\frac{4.3}{5.267}$$

.816

Table 7
Six Forms of Intraclass Correlations for Table 4 Data Set

ICC Form	Coefficient
(1,1)	.236
(2,1)	.344
(3,1)	.597
(1,3)	.481
(2,3)	.611
(3,3)	.816

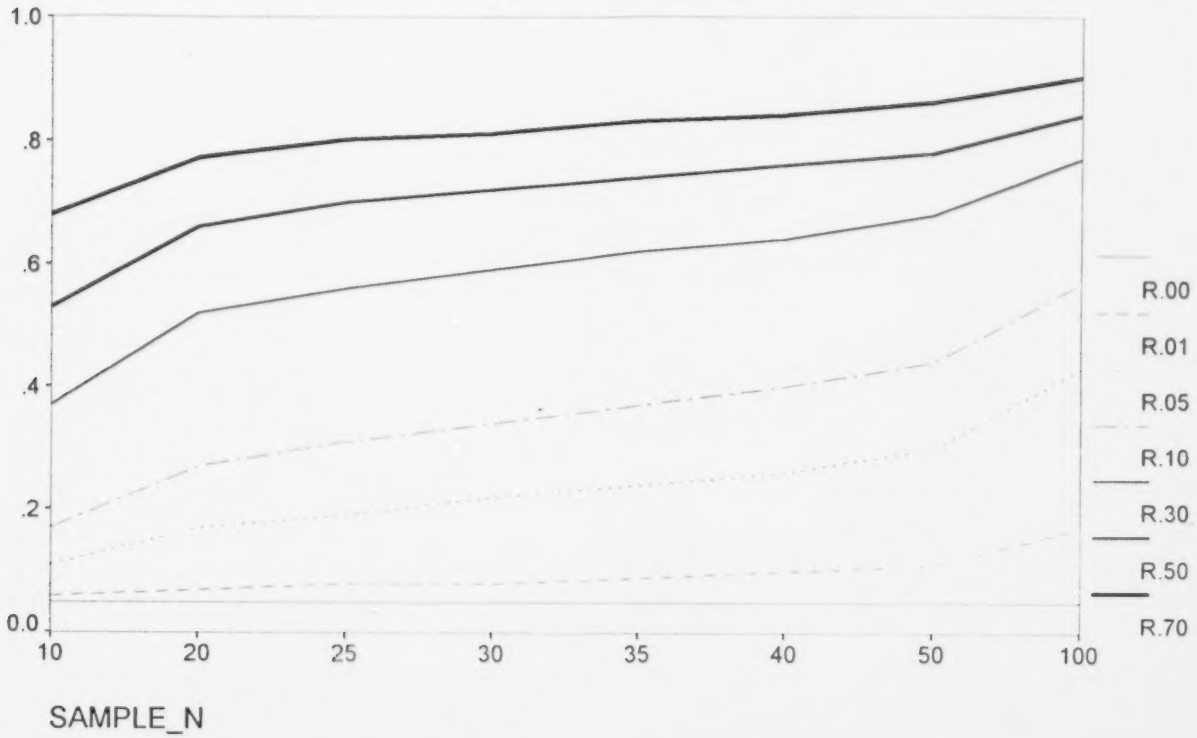

Figure 1.
True α Levels for Nominal α=.05 and n's Ranging from 10 to 100
with Intraclass Correlations Ranging from .00 to .70

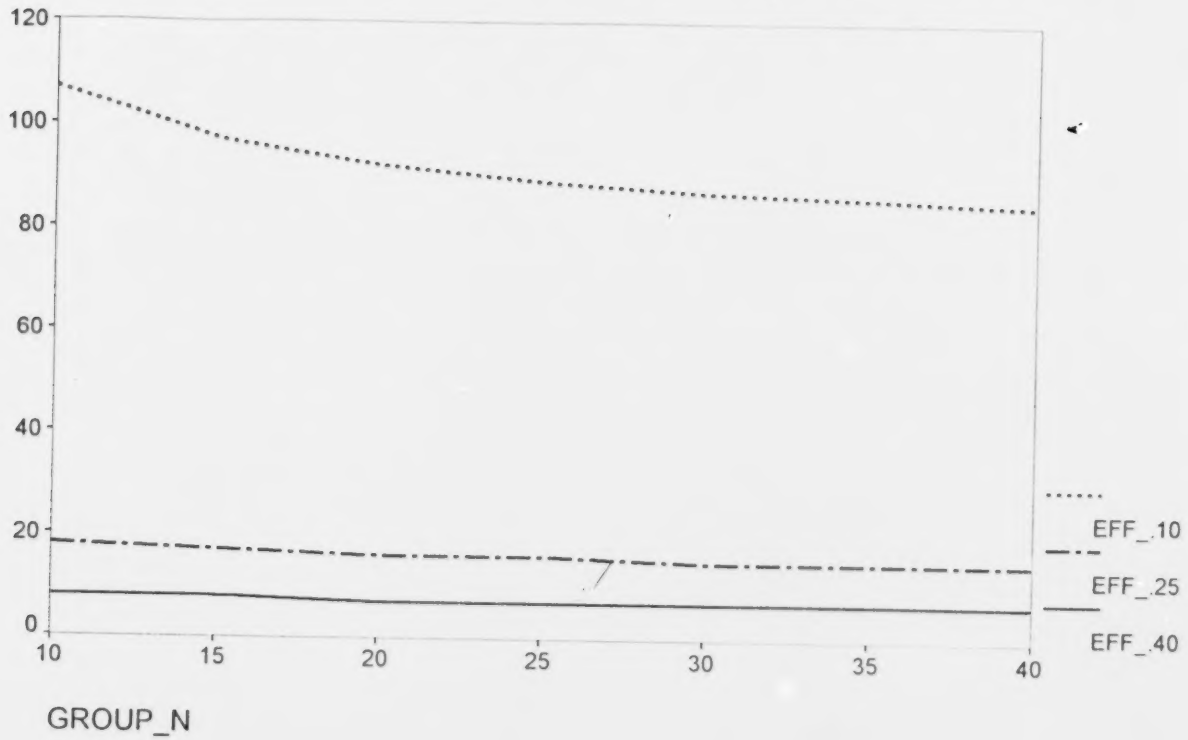

Figure 2.
Number of Groups Needed to Attain Power of at Least .80 for Groups Sizes Ranging from 10 to 40 and Effect Sizes of ¦.10¦, ¦.25¦ or ¦.40¦